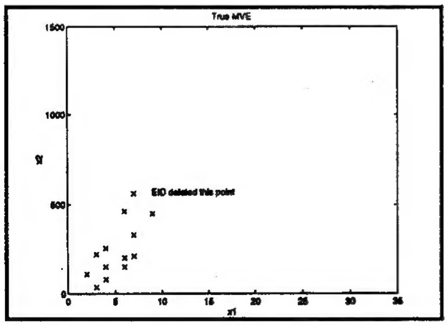
ARO 32850.1-MA

A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID

Wendy L. Poston, Edward J. Wegman,
Carey E. Priebe and Jeffrey L. Solka

Technical Report No. 106
October, 1994

Center for
Computational
Statistics



George Mason University
Fairfax, VA 22030

CENTER FOR COMPUTATIONAL STATISTICS
TECHNICAL REPORT SERIES (RECENT REPORTS)

TR 93. Winston C. Chow, Modeling and Estimation with Fractional Brownian Motion and Fractional Gaussian Noise (Ph.D. Dissertation), February, 1994.

TR 94. Mark C. Sullivan and Edward J. Wegman, Correlation Estimators Based on Simple Nonlinear Transformations, February, 1994, To appear **IEEE Transactions on Signal Processing**.

TR 95. Mark C. Sullivan and Edward J. Wegman, Normalized Correlation Estimators Based on Simple Nonlinear Transformations, March, 1994.

TR 96. Kathleen Perez-Lopez and Arun Sood, Comparison of Subband Features for Automatic Indexing of Scientific Image Databases, March, 1994.

TR 97. Wendy L. Poston and Jeffrey L. Solka, A Parallel Method to Maximize the Fisher Information Matrix, June, 1994.

TR 98. Edward J. Wegman and Charles A. Jones, Simulating a Multi-target Acoustic Array on the Intel Paragon, June, 1994.

TR 99. Barnabas Takacs, Edward J. Wegman and Harry Wechsler, Parallel Simulation of an Active Vision Model, June, 1994.

TR 100. Edward J. Wegman and Qiang Luo, Visualizing Densities, October, 1994.

TR 101. Daniel B. Carr, Converting Tables to Plots, October, 1994.

TR 102. Julia Corbin Fauntleroy and Edward J. Wegman, Parallelizing Locally-Weighted Regression, October, 1994.

TR 103. Daniel B. Carr, Color Perception, the Importance of Gray and Residuals on a Choropleth Map, October, 1994.

TR 104. David J. Marchette, Carey E. Priebe, George W. Rogers and Jeffrey L. Solka, Filtered Kernel Density Estimation, October, 1994.

TR 105. Jeffrey L. Solka, Edward J. Wegman, Carey E. Priebe, Wendy L. Poston and George W. Rogers, A Method to Determine the Structure of an Unknown Mixture Using the Akaike Information Criterion and the Bootstrap, October, 1994.

TR 106. Wendy L. Poston, Edward J. Wegman, Carey E. Priebe and Jeffrey L. Solka, A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID, October, 1994.

TR 107. Clifton D. Sutton, Tree Structured Density Estimation, October, 1994.

TR 108. Charles A. Jones, Simulating a Multi-target Acoustic Array on the Intel Paragon (M.S. Thesis), October, 1994.

A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID

Wendy L. Poston[1], Edward J. Wegman[2], Carey E. Priebe[3], & Jeffrey L. Solka[4]

[1]Naval Surface Warfare Center, Dahlgren Div, G33
Dahlgren, Virginia 22448-5000

[2]George Mason University
Fairfax, Virginia 22030

[3]Johns Hopkins University
Baltimore, Maryland 21218

[4]Naval Surface Warfare Center, Dahlgren Div, B10
Dahlgren, Virginia 22448-5000

ABSTRACT

The existence of outliers in a data set and how to deal with them is an important problem in statistics. The Minimum Volume Ellipsoid (MVE) estimator is a robust estimator of location and shape; however its use has been limited because few computationally attractive methods exist to calculate it. Determining the MVE consists of two parts: finding the subset of points to be used in the estimate and finding the ellipse that covers this set. This paper will address the first problem. The proposed method of subset selection is called the Effective Independence Distribution (EID) method which chooses the subset by mnimizing determinants of matrices containing the data. This method is deterministic yielding reproducible estimates of location and scatter for a given data set. The EID method of finding the MVE is applied to several regression data sets where the true estimate is known. Results show that the EID method produces the subset of data in less than a second and that there is less than 6% relative error in the estimates.

1. INTRODUCTION

An important part of research in statistical theory is the robust estimation of location and covariance structure for a set of data. In this paper robust estimation will refer to those estimators that have high breakdown points [Rousseeuw & Leroy, 1987]. The estimator of interest here is called the Minimum Volume Ellipsoid (MVE). This has desirable robustness properties due to its high breakdown point of 50% [Woodruff & Rocke, 1993]. Few computationally reasonable methods of determing the MVE exist, especially in high dimensions and for large sample sizes, which makes it impractical for frequent use by statisticians.

The MVE is defined as the subset of h points subject to the constraint that the ellipsoid that covers the points has minimum volume [Rousseeuw, 1985; Hawkins, 1993; Woodruff & Rocke, 1993]. As such, it is an estimator that has minimum volume and high content. From this definition of the MVE, it is apparent that finding a value of the estimator for a given data set has two parts. The first is to find the subset of data that is to be included in the estimate, and the second is to calculate the covering ellipsoid. An algorithm has been published (ref) that will find the exact covering ellipse for a set of points. However, it still requires exhaustive specification of all possible sub-samples, making it computationally intensive for large data sets. Thus, it should be apparent that the subset selection problem is the more computationally intensive of the two problems, and the one the remains to be solved. It is this issue that will be addressed in this paper.

Current methods of subset selection include the basic resampling method described by Rousseeuw and Leroy [1987] which randomly chooses subsets and then keeps the one yielding the minimum volume as the answer. Improvements on this include heuristic search algorithms investigated by Woodruff and Rocke [1993]. Yet another approach to finding the MVE is that of Hawkins [1993] called the Feasible Solution Algorithm (FSA). All of these methods are random, and they are not guaranteed to find the exact MVE for any finite amount of sampling. Clearly, none of these methods provide reproducible estimates of the MVE for a given data set.

The FSA method determines a candidate subset of h points randomly and then weights each point until all of them are covered by an ellipse of smallest volume for the current set of points. Pairwise exchanges of covered for uncovered points is then made, and the weights are re-adjusted to cover the points. If the volume decreases, then the starting set of h points will not yield the MVE. This exchange continues, and if no pairwaise exchange leads to a covering ellipsoid of smaller volume, then the candidate set of points is a feasible solution. This continues for different candidate sets of h points, and the set yielding the smallest volume is accepted as the MVE. Hawkins does propose some improvements to this basic algorithm that provide a sppedup in the computations involved.

Woodruff and Rocke [1993] investigate several heuristic search algorithms for finding the MVE. Specifically, these techniques are: simulated annealing, tabu search and genetic algorithms. In this paper, they show that the random search method is dominated by these. A basic genetic algorithm begins with many possible solutions to the MVE, and then new solutions are formed using crossover and mutation operations. One parent of a new solution is chosen based on how well it met the optimization objective and the other is chosen randomly. The crossover point (at what point the sample will be swapped) is chosen randomly, and the mutation operator looks at each value in the sample and with some low probablity changes it.

Simulated annealing is another random method that is based on steepest descent. It escapes from possible local minima by accepting with some probability a worse solution. As the process continues, there is less chance of accepting a worse answer as one would desire. Tabu search also accepts new solutions based on a steepest descent design, except that a tabu list is used to force the search away from solutions that were examined in recent iterations.

The computationally expensive part of determining the MVE is that of finding the subset of points to be covered by the ellipse. The Effective Independence Distribution (EID) method [Poston, 1994] is proposed as a solution to the subset selection problem. As with the other methods, it does not provide the exact MVE. However, results will be presented that show that it does pick subsets that yield ellipsoids approaching the true MVE. Other aspects that make it particularily appealing are the repeatability of an estimate for a given data set due to its deterministic nature, and the fact that it is computationally tractable even for large data sets and high dimensional problems.

Some background information on the MVE estimator and the EID method is provided. The algorithm for determining the weights such that all of the points are covered is described in Hawkins [1993] and is repeated here for completeness. Results are presented that show the relative error in the volume of the ellipsoid found using the EID approach for several regression data sets where the true MVE is known.

2. MINIMUM VOLUME ELLIPSOID ESTIMATOR

The problem of robust estimation of multivariate location and shape is: given a set of *n p*-dimensional observations, find an estimate of location and shape that is resistant to outliers or contaminated data. The MVE is one such estimator, and it is known that it has a breakdown point that approaches 50% as the number of points in the data set increases. This is the maximum possible breakdown point, and it means that approximately half of the data can be arbitrarily contaminated without affecting the estimate.

The MVE is given by the ellipsoid [Hawkins, 1993]

$$(\mathbf{x}-\mathbf{c})^T \Gamma^{-1} (\mathbf{x}-\mathbf{c}) = p \tag{2.1}$$

where $\mathbf{c}$ and Γ are the location vector and scatter matrix respectively and p is the dimension of the data. The location vector is a weighted mean calculated as

$$\mathbf{c} = \sum_{i=1}^{h} w_i \mathbf{x}_i^* \tag{2.2}$$

and the covariance or scatter matrix is

$$\Gamma = \sum_{i=1}^{h} w_i (\mathbf{x}_i^* - \mathbf{c})(\mathbf{x}_i^* - \mathbf{c})^T \tag{2.3}$$

where $\mathbf{x}_i^*$ is a column vector denoting the *ith* observation in the subset of h points, w_i is the weight for the *ith* observation, and $h = [(n+p+1)/2]$ (the brackets denote the greatest integer function). The volume of the covering ellipse will be proportional to the determinant of Γ. It is evident from these equations that to find the MVE one must determine which h points should be covered and the corresponding weights to ensure coverage of the points.

The algorithm that will be used to find the weights is credited to Titterington [1975] and is described in Hawkins [1993]. It will be referred to in the sequel as Titterington's algorithm. All of the weights are initially set to $w_i^{(0)} = 1/h, \quad i = 1, \ldots, h$, which is the usual weight given to points when calculating the sample mean of a data set of size h. Then at each iteration k, calculate the weighted mean and covariance from Eqs. (2.2)-(2.3) and the Mahalanobis distances for each observation given by

$$D_i^{(k)} = (\mathbf{x}_i^* - \mathbf{c}^{(k)})^T \Gamma^{-1} (\mathbf{x}_i^* - \mathbf{c}^{(k)}) \tag{2.4}$$

If $D_i^{(k)} \le p$ for every i, then the current ellipsoid using $\mathbf{c}^{(k)}$ and $\Gamma_{(k)}^{-1}$ is the MVE covering the h observations. If the Mahalanobis distance for any of the observations exceeds p, then the weights must be adjusted using the following

$$w_i^{(k+1)} = w_i^{(k)} \frac{D_i^{(k)}}{p} \tag{2.5}$$

and the calculations of Eqs. (2.2)-(2.4) are repeated until all of the distances are less than p. This procedure enlarges the ellipsoid until all of the h points are covered.

The algorithm for finding the weights can be somewhat computationally intensive for some data sets. However, it should be apparent that the real computational burden arises from the determination of which points should be covered by the ellipse. The EID algorithm is presented as a means of addressing this problem.

3. EFFECTIVE INDEPENDENCE DISTRIBUTION

3.1 The Development of EID from an Eigenvalue Problem

The derivation of the EID provided here was first given by Kammer [1991]. Subsequent research has shown a similar idea proposed by Rousseeuw and Leroy [1987]. They proposed using the diagonal elements of the 'hat' matrix to remove outliers from the data set, which of course is the purpose also of the MVE. The EID provides a ranking of each point according to its contribution to the eigenvalues, and hence to the determinant, of the FIM. It will be shown that the EID offers a direct relationship between the determinants of the information matrix as points are removed from the data set. Thus, the EID used in the method described here optimizes the determinant of the Fisher Information Matrix (FIM), which is defined below.

The EID is developed from the set of equations familiar from regression theory [Rousseeuw & Leroy, 1987]. These are

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{3.1}$$

where **y** is an *n*-dimensional vector of responses, **X** is an *n x p* matrix of predictor variables, β is a *p*- dimensional column vector of unobservable parameters that must be estimated from the data, and ε denotes the noise in the measurements. It is further assumed that

$$E[\varepsilon] = 0$$

and

$$E[(\varepsilon - \mu)^2] = \Sigma$$

Without loss of generality, the covariance matrix of the noise, Σ, is assumed to be the identity matrix. Thus, the FIM is given by

$$FIM = \mathbf{X}^T\mathbf{X} \tag{3.2}$$

and the covariance matrix providing statistical information for the estimate of β is the inverse of the FIM.

To get a good estimate of the parameters β, measurements should be chosen that will minimize a norm on the covariance matrix. Alternatively, one could maximize a norm on the inverse of the covariance matrix. Thus, the objective function will be the determinant of the FIM, and it will be the goal of this derivation to show that the EID can be used to delete points in such a way that the determinant is optimized.

The EID is an n-dimensional vector where each element corresponds to one measurement location. The development of the EID method given here will show that the *ith* term of the EID vector is the contribution of the *ith* measurement to all of the eigenvalues of the FIM. Since

$$|FIM| = \prod_{j=1}^{p} \lambda_j \tag{3.3}$$

where $(|\bullet|)$ denotes the determinant, then the eigenvalues are also a measure of the information and indicate the contribution of a measurement to the determinant of the FIM.

The EID can be derived from the following eigenvalue problem

$$(FIM - \lambda_j \mathbf{I})\Psi_j = 0 \tag{3.4}$$

where $\mathbf{I}$ is a $p \ x \ p$ identity matrix, λ_j is the *jth* eigenvalue, and Ψ_j is the *jth* eigenvector. It follows from the definition of the information matrix that the FIM is symmetric. Since the columns of $\mathbf{X}$ are linearly independent, this implies that the FIM is positive definite. Therefore, the eigenvector Ψ_j can be chosen to be orthonormal, [Strang, 1988] which implies that

$$\Psi_i^T \Psi_j = 0, \quad i \neq j$$

and

$$\Psi_i^T \Psi_i = 1$$

Hence, the following matrix properties hold

$$\begin{aligned} \Psi^T \Psi &= \mathbf{I} \\ FIM\,\Psi &= \Lambda \Psi \end{aligned} \tag{3.5}$$

where Ψ is an orthonormal matrix with each column containing an eigenvector and Λ denotes a diagonal matrix of eigenvalues.

Starting from the second property given above and substituting for the FIM, yields

$$\mathbf{X}^T\mathbf{X}\Psi = \Lambda\Psi \tag{3.6}$$

Pre-multiplying by Ψ^T and using the first property in Eq. (3.5 gives

$$\Psi^T\mathbf{X}^T\mathbf{X}\Psi = \Lambda \tag{3.7}$$

After grouping terms, this can be written as

$$(\mathbf{X}\Psi)^T(\mathbf{X}\Psi) = \Lambda \tag{3.8}$$

It can be seen from this that the *jth* eigenvalue has the form

$$\lambda_j = \sum_{i=1}^{n}\left(\sum_{j=1}^{p}\mathbf{x}_{ik}\psi_{kj}\right)^2, \quad j = 1,\ldots,p \tag{3.9}$$

The eigenvectors of the information matrix span the *p*-dimensional parameter space, so Ψ can be used to transform the data matrix **X**. The following matrix product is now formed

$$\mathbf{G} = (\mathbf{X}\Psi)\otimes(\mathbf{X}\Psi) \tag{3.10}$$

where $\otimes$ denotes an element by element matrix multiplication and $\mathbf{X}\Psi$ represents the transformed mode shape matrix. The *ij-th* element of **G** is given by

$$g_{ij} = \left(\sum_{k=1}^{p}\mathbf{x}_{ik}\psi_{kj}\right)^2 \tag{3.11}$$

An examination of each element of **G** reveals that the sum of the *jth* column of **G** equals the *jth* eigenvalue given in Eq. (3.9).

$$\sum_{i=1}^{n} g_{ij} = \lambda_j \tag{3.12}$$

The next step is to post-multiply **G** by Λ^{-1} forming the following matrix

$$\mathbf{E} = \mathbf{G}^{-1}\Lambda \tag{3.13}$$

The purpose of this step is to normalize each column of **G** by dividing by the corresponding eigenvalue (i.e., the *jth* column is divided by the *jth* eigenvalue). Each column in the matrix **E** sums to one, and the element e_{ij} represents the fractional contribution of the *ith* measurement or data point to the *jth* eigenvalue.

The EID is calculated by summing the terms in the *ith* row of the matrix **E**

$$EID_i = \sum_{j=1}^{p} e_{ij} \tag{3.14}$$

Thus, EID$_i$ represents the contribution of the *ith* observation to the eigenvalues of the FIM. Again, note that there are *n* elements in the EID, one corresponding to each point in the data set.

3.2 An Alternative Calculation of the EID

The diagonal elements of the following matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{3.15}$$

will also yield the EID values for each observation. To derive this equation, start with the definition of the *ith* element of the EID

$$EID_i = \sum_{j=1}^{p} e_{ij} = \sum_{j=1}^{p} \frac{g_{ij}}{\lambda_j} \tag{3.16}$$

and substituting for the *ij-th* element of **G** from Eq. (3.11) yields

$$EID_i = \sum_{j=1}^{p} \left(\sum_{k=1}^{p} \frac{\mathbf{x}_{ik}\psi_{kj}}{\sqrt{\lambda_j}} \right)^2 \tag{3.17}$$

These are the diagonal elements of the following matrix product

$$\mathbf{H} = (\mathbf{X}\Psi\Lambda^{-1/2})(\mathbf{X}\Psi\Lambda^{-1/2})^T \tag{3.18}$$

where $\Lambda^{1/2}$ is a diagonal matrix containing the square roots of the eigenvalues. Rearranging the matrices yields

$$\mathbf{H} = \mathbf{X}\Psi\Lambda^{-1}\Psi^T\mathbf{X}^T \tag{3.19}$$

Using the properties in Eq. (3.5), it can be shown that

$$FIM^{-1} = \Psi\Lambda\Psi^{T} \tag{3.20}$$

Thus, substituting Eq. (3.2) and Eq. (3.20) into Eq. (3.19), the matrix **H** can be re-written as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T} \tag{3.15}$$

and the elements of the EID can be calculated from

$$EID = diag(\mathbf{H}) \tag{3.21}$$

The matrix given in Eq. (3.15) is the familiar 'hat' matrix from regression theory [Rousseeuw & Leroy, 1987]. It has some interesting properties that offer some insight into the nature of the EID. One is that it is an idempotent matrix. These matrices have the proerty that the trace equals the rank, so

$$\sum_{i=1}^{n} EID_{i} = rank(\mathbf{H}) = rank(\mathbf{X}) = p \tag{3.22}$$

The EID can be said to show the contribution of the *ith* measurement location to the rank of the data matrix and thus also to the linear independence of the parameter space.

3.3 Motivation for Using the EID for Subset Selection

It has been shown previously [Poston & Tolson, 1992] that the following relationship holds between the determinants of the FIM as points are removed from a data set

$$\left|\mathbf{X}_{-i}^{T}\mathbf{X}_{-i}\right| = (1 - EID_{i})\left|\mathbf{X}^{T}\mathbf{X}\right| \tag{3.23}$$

where $\mathbf{X}_{-i}$ is the data matrix with the *ith* point removed and $\mathrm{EID_i}$ is the value for the *ith* point. From this one can see that there is a direct relationship between the determinants as the points are removed from the data set. If the objective is to minimize the determinant, then the observation with the largest EID value should be deleted. This is the case for finding the set pf points used to determine the MVE.

The following proposition will be proven [Rousseeuw & Leroy, 1987] about the possible range of values that an element of the EID$_i$ can have.

PROPOSITION: EID$_i$ is in the range

$$0 \leq EID_i \leq 1$$

PROOF: Since **H** is an idempotent matrix, this implies that

$$h_{ii} = (\mathbf{HH})_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji}$$

Since **H** is also symmetric, the diagonal elements can be written

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = \sum_{j=1}^{n} h_{ij}^2$$

Expanding the sum on the right-hand side yields

$$h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2$$

This equality can only be true if $h_{ii} \leq h_{ii}^2$ which implies that

$$0 \leq h_{ii} \leq 1$$

or that

$$0 \leq EID_i \leq 1$$

and the proposition is proved.

It is instructive to examine what happens if an observation has one of the extreme values of zero or one. A data point with a value of one must retained to preserve the linear independence of the data matrix **X**. This is obvious from Eq. (3.23). If such a point is deleted, then the determinant of the FIM is zero and the problem becomes singular. In the regression setting, this means that all of the parameters β cannot be estimated. On the other hand, if an observation has a value of zero, then the determinant is unchanged and no loss of information occurs.

Recall that the volume of the MVE is proportional to the determinant of Γ. This is the reationale for using the EID to select the subset of data points that is used in the MVE. If we use the matrix $\mathbf{X}^T\mathbf{X}$ to approximate the scatter matrix Γ, then we can use the relationship in Eq. (3.22) to successively remove points until the desired h points remain. These h points will then be used in the algorithm described previously for finding the weights and the resulting ellipsoid. However, to better approximate the scatter matrix, the data will be centered by subtracting the p-dimensional sample mean from each observation. This is repeated as each point is deleted. The complete procedure consists of the following steps:

1. Calculate the matrix

$$\mathbf{X}'^{(j)} = (\mathbf{X}^{(j)} - \overline{\mathbf{X}}^{(j)})$$

where $\mathbf{X}^{(j)}$ is the set of raw data points at the *jth* iteration of the method and $\overline{\mathbf{X}}^{(j)}$is an *(n-j) x p* matrix with each row containing the p-dimensional sample mean for the current set of data. Note that at iteration j=0 there are n points in the data set, at iteration j=1 there are n-1 points, etcetera.

2. Use the matrix $\mathbf{X}'^{(j)}$ in Eq. (3.21) to calculate the EID value for each point in the current data set.
3. Delete the point that corresponds to the maximum EID value.
4. Repeat steps 1-3 until h points remain.
5. Adjust the weights using Titterington's algorithm until the h points are covered by the ellipse.

Some care should be taken with step 3 when implementing this method. It is quite possible that in the very first calculation of the n EID values that a data point has a value of one. Such a point must be kept to keep the problem nonsingular (see Eq. 3.23). Instead of deleting this point, one could remove the observation corresponding to the next highest EID value. How this affects the estimate of the MVE is a topic of ongoing research. The chances of an observation having an EID value of one becomes greater as the data set is reduced, and it is obvious from Eq. (3.22) that when there are only p points left in the set, then each observation must have a value of one. Thus, this discussion becomes more critical as more points are deleted from the set.

IV. APPLICATIONS AND RESULTS

To test the usefulness of this method, it is applied to several data sets where the true MVE is known. The paper by Hawkins [1993] gives the correct subset and the resulting volume of the true MVE for these data sets. The relative error in the volume of the ellipse based on the subset obtained using the EID method can then be determined for comparison purpose. The 6 data sets can be found in Rousseeuw and Leroy [1987]. These data are used for regression purposes, and only the predictors are used here to determine the MVE. The parameters of interest are shown in Table I. From this one can see that the data sizes are relatively small ranging in size from n=20 to n=50. The dimensionality of the data is also low, from 2 to 5 dimensions.

For this study, the EID algorithm is implemented in MATLAB on a 486, 33MHz computer. The relative error in the volumes of the minimum covering ellipsoid using the EID approach is shown in Figure 1. It is evident from the small error that ours is a feasible approach to finding the MVE.

The time needed to determine the subset of points is given in Table II. Also in this table are some timings obtained using Splus 3.1 to determine the MVE estimate of a covariance matrix. This software uses a genetic algorithm to find the subset of points. These results are presented to provide a very rough comparison of the two methods in terms of the computational effort involved. One can see that using the EID yields a savings in time when calculating the MVE, which would become more important as the dimensionality and size of the data set increases.

The 2-dimensional 'delivery' data set is shown in Figure 2 to provide a qualitative assessment of the method. From this, it is clear that the bulk of the data is clustered toward the the origin. When the EID method is applied to this data set, the first observations that are deleted are the outlying ones in the upper right-hand corner of the plot. It is not until the last points are deleted that the EID algorithm makes an incorrect choice. The set chosen by the EID approach is shown in Figure 4. Note the point that is incorrectly retained in the set. One reason for this error is that the point the EID deletes has a larger magnitude than the one that should be kept in the set. Previous studies indicate that these will be the points that tend to have a large EID value

Finally, one last comparison is in order regarding the 'salinity' data set. It is stated in Hawkins [1987] that this set would require approximately 5,000 random starts with the FSA to reliably determine the MVE, which is a computationally intensive task. Note that for this data set the EID method of subset selection finds a set of points in 0.22 sec with only 3% error in the volume of the ellipse. Thus, the EID method achieves a good estimate of the true MVE even for data sets where other methods have trouble.

V. SUMMARY

In this paper, the EID method of determining the subset of points used in the MVE has been described. Subset selection is what makes the MVE a computationally expensive method to implement in daily practice. Preliminary results indicate that the EID method for selecting the set of points to be included in the MVE estimator is a useful one. The time required for subset selection is less than a second for the data sets considered here, and it is expected that for large *n* similar savings in time can be achieved.

The 2-dimensional scatterplots of the 'delivery' data indicate qualitatively that the EID tends to pick a tighter cluster of points. Whereas the set of points making up the true MVE is somewhat narrower, yielding a smaller ellipse. This example helps illustrate an important point about the MVE. Since it is an ellipsoid of minimum volume, it does not necessarily pick the tightest cluster of data. It is suspected that the EID approach might yield better results based on some other criterion; e.g., better covariance structure or clustering. These ideas are part of the ongoing research in using the EID for the robust estimation of multivariate location and scatter.

Although the EID method is not guaranteed to find the true MVE, it has certain advantages that make it more attractive than the algorithms currenly in use. As discussed previously, it involves little computational effort, and thus it is suitable for sets with large *n* and *p*. Also, due to the iterative nature of the method, it would be easy to get a family of estimators for different values of *h* which is a useful feature [Hawkins, 1993].

Since this is a deterministic method, the results are repeatable for a given data set. This is a desirable property, because one would like to achieve the same estimate of location and scatter for the same data set. Clearly, the methods currently in use are based on heuristic random searches, and thus they are not guaranteed to produce the same estimate. This feature coupled with the small computational cost makes EID the method of choice in determining the subset of points to be used in the MVE estimate of location and shape.

V. REFERENCES

Cook, R. D., Hawkins, D. M., and Weisberg, S., (1993), "Exact Iterative Computation of the robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics & Probability Letters,* p 13.

Hawkins, D. M., (1993), "A Feasible Soltion for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics,* p 95.

Kammer, D. C., (1991), "Sensor Placement for On-orbit Modal Identification and Correlation of Large Space Structures," *AIAA Journal of Guidance, Control and Dynamics,* p 251.

Poston, W. L. and Priebe, C. E., (1994), "Finding the Minimum Volume Ellipsoid," *Proceedings of the Interface, '94.*

Poston, W. L., and Tolson, R. H., (1992), "Maximizing the Determinant of the Information Matrix With the Effective Independence Distribution Method," *AIAA Journal of Guidance, Control and Dynamics,* p 1513.

Rousseeuw, P. J., (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications, Volme B,* eds. W. Grossmann, G. Pflug, I. Vincze, and W. Werz, Dordrecht: Reidel, 283-297.

Rousseeuw, P. J., and Leroy, A. M., (1987), *Robust Regression and outlier Detection,* New York, NY, John Wiley & Sons.

Titterington, D. M., (1975), "Optimal Design: Some Geometrical Aspects of D-optimality," *Biometrika,* p 313.

Woodruff, D. L. and Rocke, D. M., (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics,* p 69.

Table I. Regression Data Set Parameters

Data Set	*p*	*n*	*h*
Aircraft	4	23	14
Coleman	5	20	13
Delivery	2	25	14
Education	3	50	27
Gravity	5	20	13
Salinity	3	28	16

Table II. Timing (sec) Results for Methods to Pick the MVE

Data Set	EID	Splus, Genetic Algorithm
Aircraft	0.22	68.0
Coleman	0.17	67.0
Delivery	0.11	28.0
Education	0.77	74.0
Gravity	0.11	62.0
Salinity	0.22	50.0

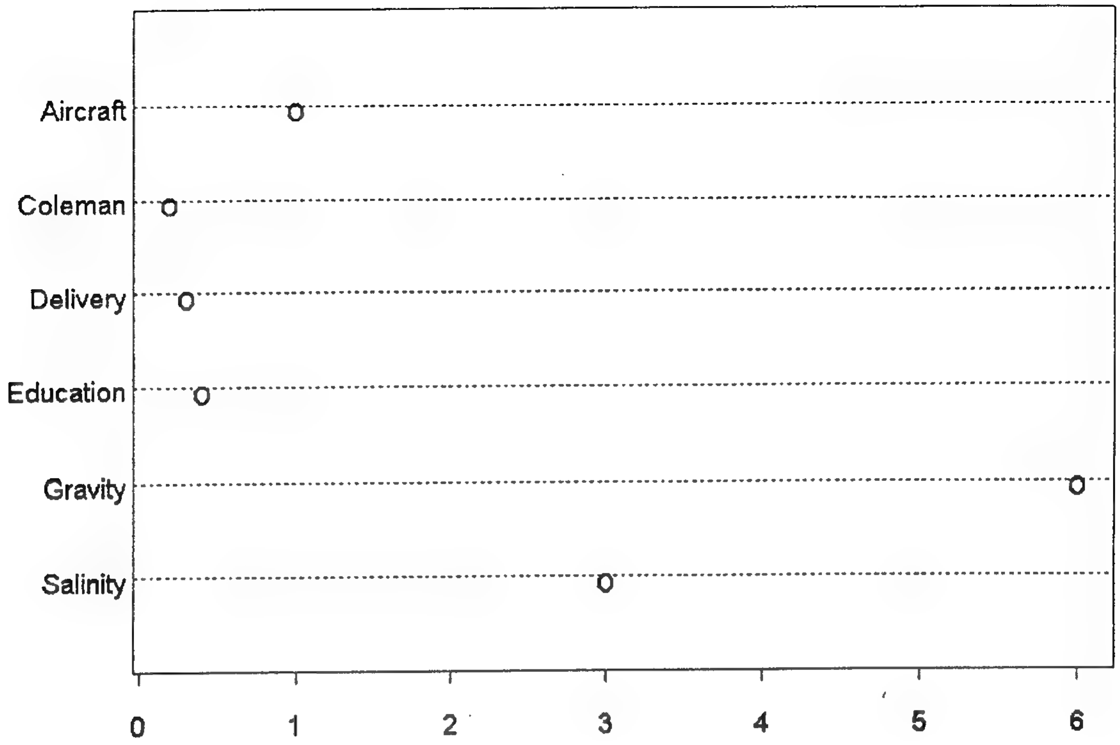


Figure 1. Percent relative error in the volume of the MVE as determined by the EID approach.

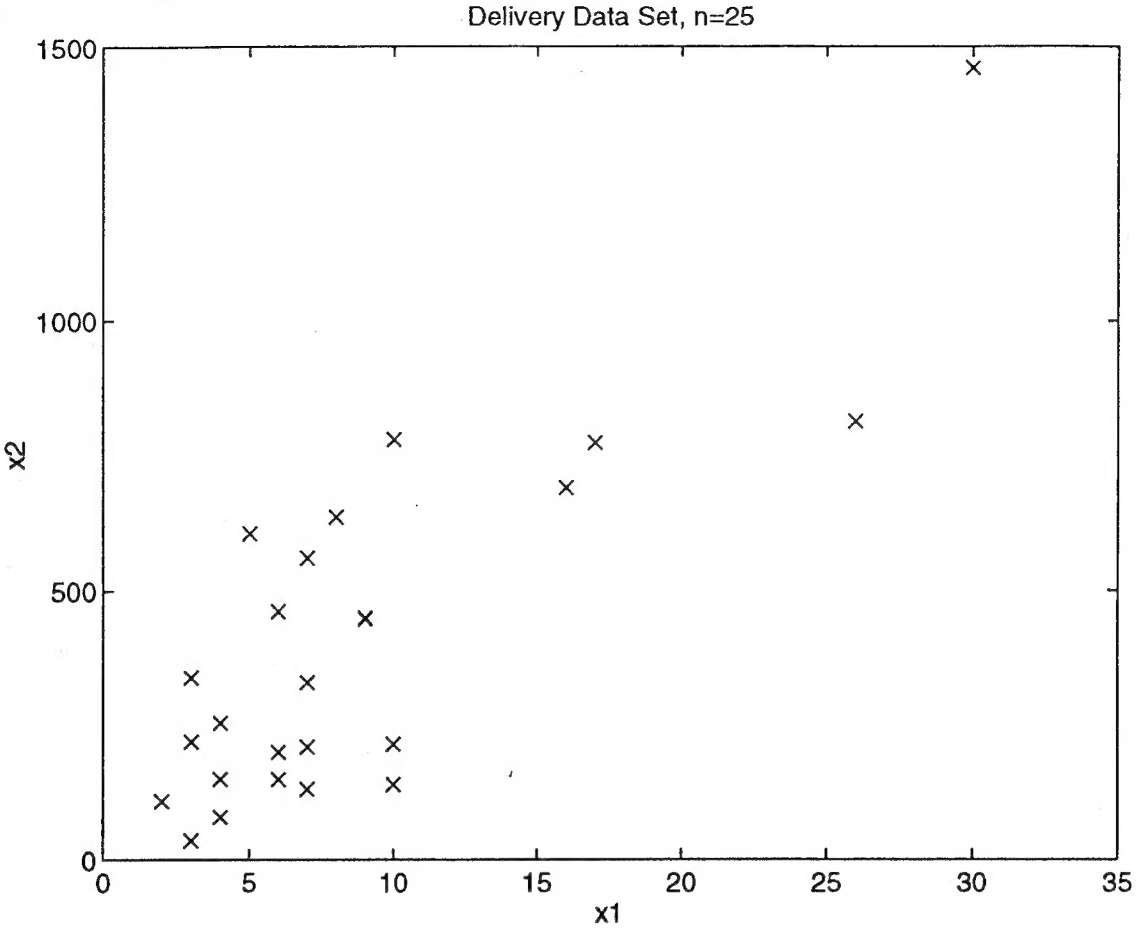


Figure 2. Scatterplot of entire 'delivery' data set. Note that most of the data is clumped near the origin.

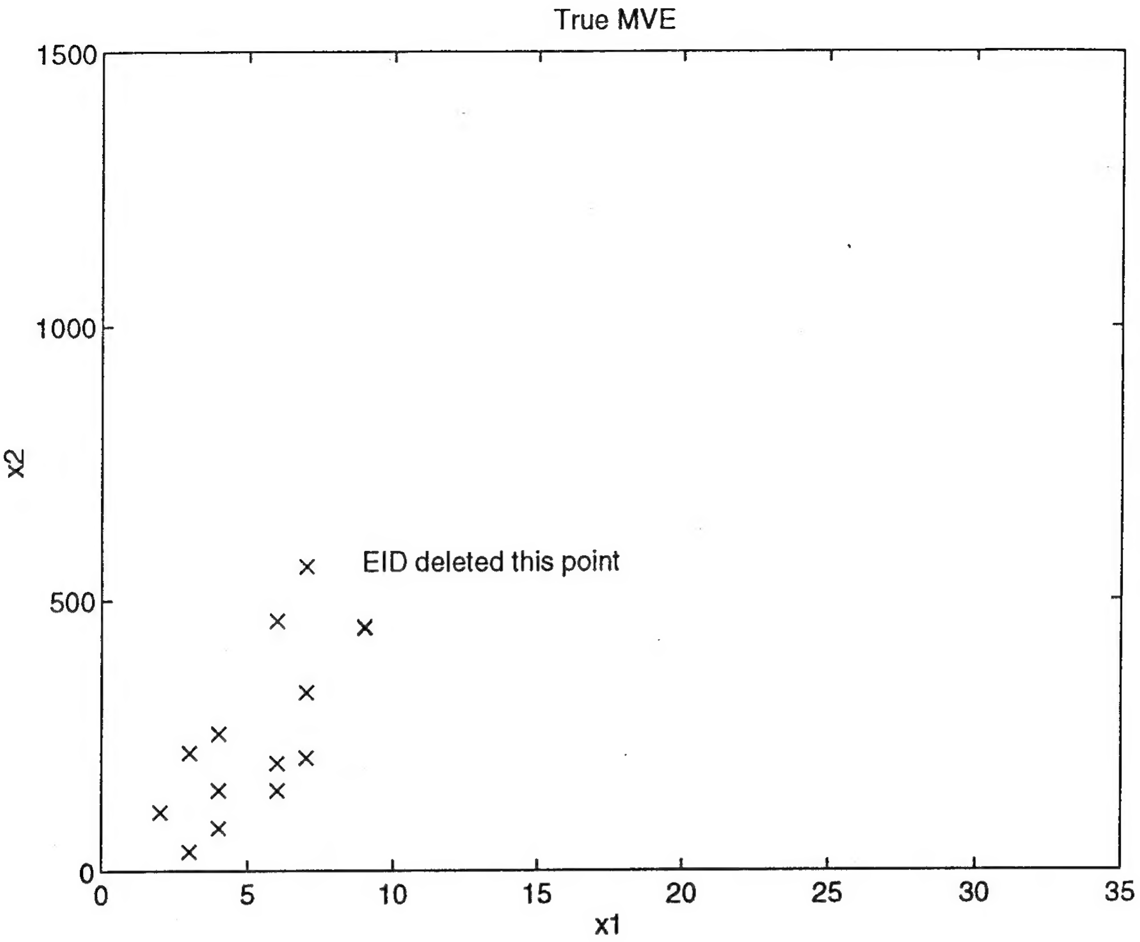


Figure 3. Scatterplot of the *h* points that are covered by the true MVE. Note the point that is incorrectly deleted by the EID algorithm.

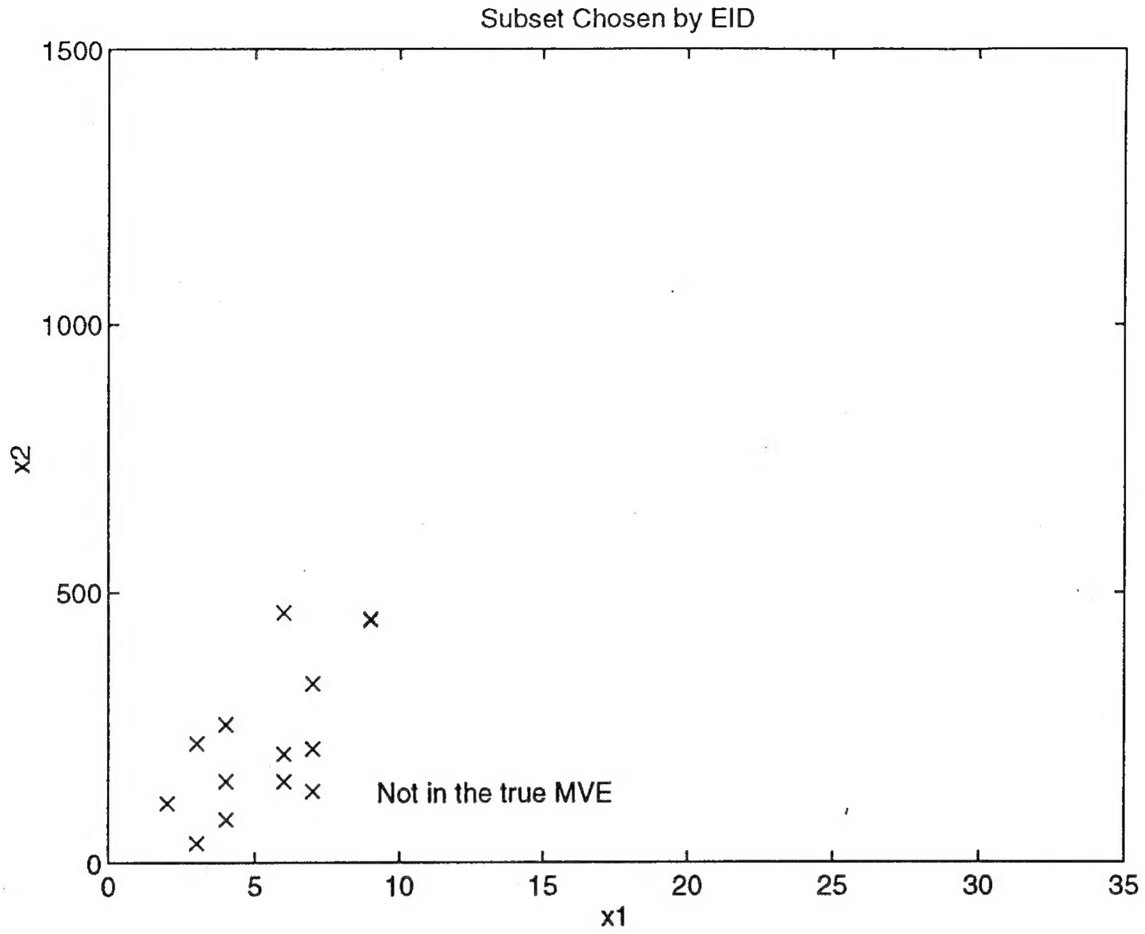


Figure 4. Scatterplot of the *h* points chosen by the EID method. Note the point that should have been deleted.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE October, 1994	3. REPORT TYPE AND DATES COVERED Technical

4. TITLE AND SUBTITLE

A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID

5. FUNDING NUMBERS

N00014-92-J-1303
N00014-93-1-0527

6. AUTHOR(S)

Wendy L. Poston, Edward J. Wegman,
Carey E. Priebe and Jeffrey L. Solka

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Computational Statistics
George Mason University
Fairfax, VA 22030

8. PERFORMING ORGANIZATION REPORT NUMBER

TR no.106

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of the Navy
Office of the Chief of Naval Research
Mathematical Sciences Division
800 N. Quincy Street Code 1111SP
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The existance of outliers in a data set and how to deal with them us an important problem in statistics. The Minimum Volume Ellipsoid (MVE) estimator is a robust estimator of location and shape; however its use has been limited because few computationally attractive methods exist to calculate it.Determining the MVE consists of two parts: finding the subset of points to be used in the estimate and finding the ellipse that covers this set.This paper will address the first problem. The proposed method of subset selection is called Effective Independence Distribution (EID) method which chooses the subset by minimizing determinants of matrices containing the data. This method is deterministic yielding reproducable estimates of location and scatter for given data set. The EID method of finding the MVE is applied to several regression data sets where the true estimate is known. Results show that the EID method produces the subset of data in less than second and that there is less than 6% relative error in the estimates.

14. SUBJECT TERMS

minimum volume ellipsoid, subset selection, clustering, random search

15. NUMBER OF PAGES

24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL